

Figure 1

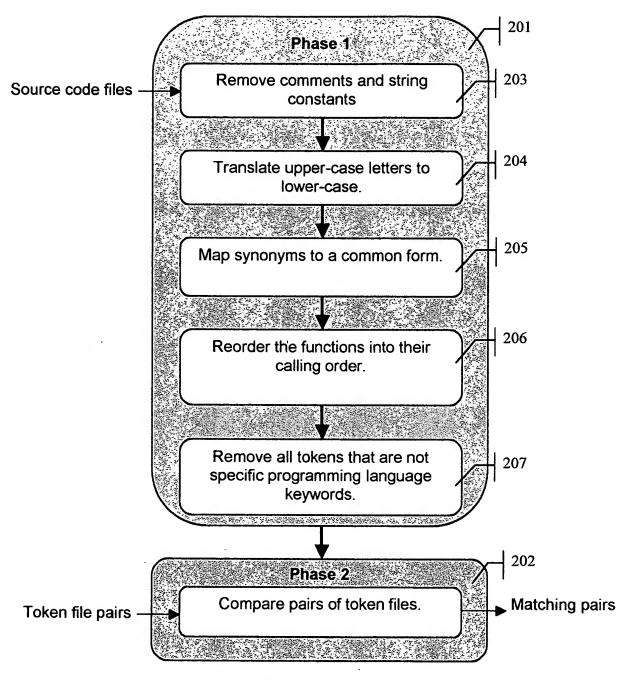


Figure 2

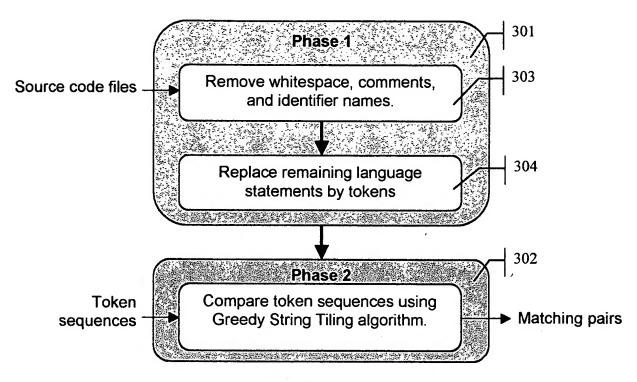


Figure 3

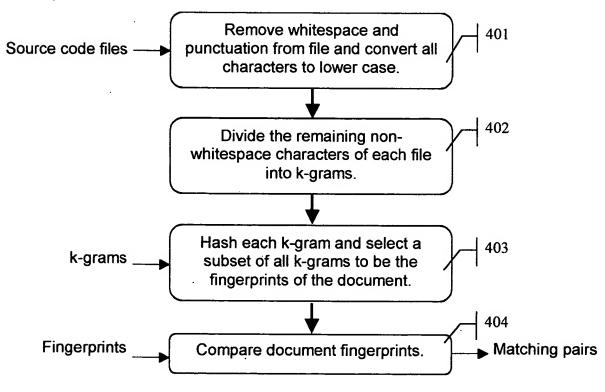


Figure 4

```
She loves you yeah, yeah, yeah.

(a) Some text.

shelo helov elove loves ovesy vesyo esyou syouy youye ouyea uyeah yeahy eahyea hyeah yeahy eahye ahyeah hyeah

(b) The sequence of 5-grams derived from the text.

502

77 72 42 17 98 50 23 55 6 66 34 24 39 11 84 24 39 11 84

(c) A hypothetical sequence of hashes of the 5-gram.

503

72 24 84 24 84

(d) The fingerprints – selecting only those hashes that are 0 mod 4.
```

Figure 5

```
---- begin
                     routine ----
void fdiv(
                           // file name
             *fname,
    char
    char *path)
                           // path
    int Index1, j;
                                                  601
    while (1)
         j =
                strlen(fname);
     /* find the file
                            extension */
                                                              602
(a) C source code snippet for file 1.
SourceLines1[0] = ""
                                             CommentLines1[0] = "---- begin routine ----"
SourceLines1[1] = "void fdiv"
                                             CommentLines1[1] = ""
SourceLines1[2] = "char fname"
                                             CommentLines1[2] = "file name"
SourceLines1[3] = "char path"
                                             CommentLines1[3] = "path"
SourceLines1[4] = ""
                                             CommentLines1[4] = ""
                                             CommentLines1[5] = ""
SourceLines1[5] = "int Index1 j"
SourceLines1[6] = ""
                                             CommentLines1[6] = ""
SourceLines1[7] = "while 1"
                                             CommentLines1[7] = ""
SourceLines1[8] = "j strlen fname"
                                             CommentLines1[8] = ""
SourceLines1[9] = ""
                                             CommentLines1[9] = "find the file extension"
(b) Source code and comment line arrays for file 1.
Word1[0] = "fdiv"
Word1[1] = "fname"
Word1[2] = "path"
Word1[3] = "Index1"
(c) Array of unique identifiers (non-keywords) in file 1.
```

Figure 6

```
Word1[0] = "abc"
                            Word2[0] = "Aabc"
                            Word2[1] = "aBc"
Wordl[1] = "abc1"
Word1[2] = "abc123"
                            Word2[2] = "abc1111111"
Word1[3] = "abcdef"
                            Word2[3] = "abcXXXyz"
Word1[4] = "pdq"
                            Word2[4] = "i"
Word1[5] = "xxx"
                            Word2[5] = "j"
Word1[6] = "xyz"
                            Word2[6] = "pdq"
Word1[7] = "yyy"
                            Word2[7] = "X"
(a) Non-keyword words in files 1 and 2.
                                                   701
PartialWord[0] = "abc"
PartialWord[1] = "abc1"
PartialWord[2] = "xxx"
PartialWord[3] = "xyz"
(b) Matching partial words
                                                   702
```

Figure 7

```
File 1
                                                         File 2
1
    /* ---- begin routine ---- */
                                             /* find the file extension */
                                       1
2
    void fdiv(
                                       2
                                             void file_divide(
3
        char *fname, // file name
                                                       *fname,
                                       3
                                             char
4
        char *path)
                     /* path */
                                             char
                                                       *path)
5
6
        int Index1, j;
                                       6
                                               int i, j;
7
                                       7
                                               while (1)
                                                             // loop here
        while (1)
                                       8
                                                j = strlen(fname);
           j = strlen(fname);
                                       9
   // find the file extension
                                       10
(a) Two files.
                                                    801
              3/3
              4/4
              9/8
                                                    802
(b) Matching source lines in file1/file2
```

Figure 8

```
File 1
                                                                    File 2
    /* ---- begin routine ---- */
                                                /* find the file extension */
                                           1
2
    void fdiv(
                                           2
                                                void file divide(
3
         char *fname, // file name
char *path) /* path */
                                                char
                                                           *fname,
4
                                                           *path)
                                                                       // path
                                                char
5
                                                {
                                                  int i, j; /* ---- begin routine ---- */
while (1) // loop here
6
         int Index1, j;
7
                                           7
         while (1)
                                           8
                                                    j = strlen(fname);
             j = strlen(fname);
                                           9
10
         // find the file extension
                                           10
                                                  switch (x)
         if (x == 5) {
11
                                                  {
(a) Two files.
               1/6
               4/4
               10/1
                                                                  902
(b) Matching comment lines in file1/file2
```

Figure 9

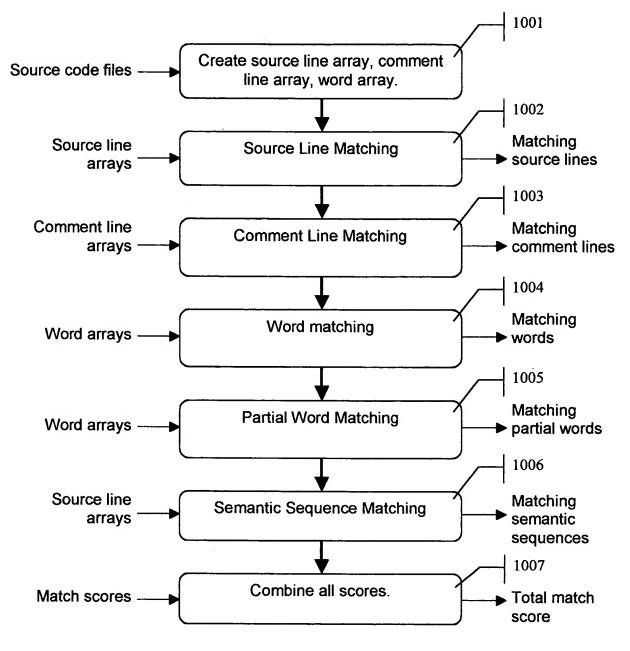


Figure 10

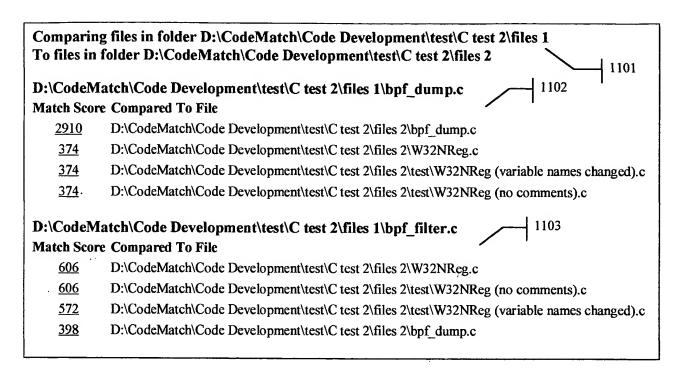


Figure 11

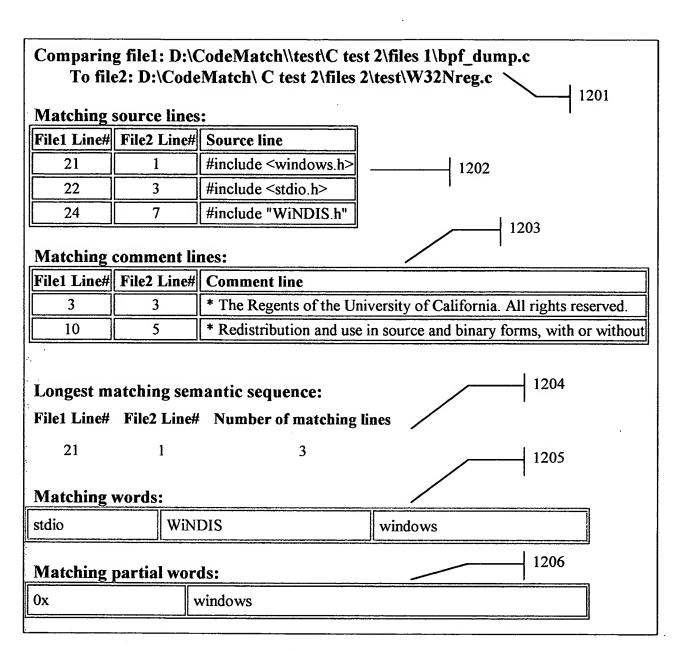


Figure 12